Stanford



Tobias Gerstenberg

Associate Professor of Psychology and, by courtesy, of Computer Science

Curriculum Vitae available Online

CONTACT INFORMATION

- **Alternate Contact**

  Reneé Rittler - Administrative Services Manager

  Email rittler@stanford.edu

  Tel 6507237431

Bio

ACADEMIC APPOINTMENTS

- Associate Professor, Psychology
- Associate Professor (By courtesy), Computer Science
- Member, Bio-X
- Faculty Affiliate, Institute for Human-Centered Artificial Intelligence (HAI)
- Member, Wu Tsai Neurosciences Institute

ADMINISTRATIVE APPOINTMENTS

- Assistant Professor of Cognitive Psychology, Stanford University, (2018- present)
- Postdoctoral Associate, Massachusetts Institute of Technology, (2014-2018)
- Postdoctoral Fellow, Massachusetts Institute of Technology, (2013-2014)

HONORS AND AWARDS

- Stanton Prize, Society for Philosophy and Psychology (2024)

PROGRAM AFFILIATIONS

- Symbolic Systems Program

PROFESSIONAL EDUCATION

- PhD, University College London , Cognitive Science (2013)
- MSc, University College London , Cognitive and Decision Sciences (2008)
- Vordiplom, Humboldt University Berlin , Psychology (2007)

LINKS

- Causality in Cognition Lab: http://cicl.stanford.edu/

Teaching

COURSES

2026-27

- Senior Honors Research: PSYCH 198 (Win, Spr)

2025-26

- Senior Honors Research: PSYCH 198 (Aut, Win, Spr)
- Statistical Methods for Behavioral and Social Sciences: COMM 352, PSYCH 252 (Win)

2024-25

- Senior Honors Research: PSYCH 198 (Aut, Win, Spr)
- Statistical Methods for Behavioral and Social Sciences: COMM 352, PSYCH 252 (Win)

2023-24

- Senior Honors Research: PSYCH 198 (Win, Spr)
- Statistical Methods for Behavioral and Social Sciences: PSYCH 252 (Win)

STANFORD ADVISEES

Doctoral Dissertation Reader (AC)

Alex Durango, Jacqueline Harding, Lynde Wolf-Gorneault

Postdoctoral Faculty Sponsor

Erik Brockbank, Yang Xiang

Postdoctoral Research Mentor

Dae Houlihan

Doctoral (Program)

Justin Yang

Publications

PUBLICATIONS

- **A Communication-First Account of Explanation** *NOUS*
  Harding, J., Gerstenberg, T., Icard, T.
  2026
- **Seeing, hearing, and feeling causation.** *Cognitive psychology*
  Chase, E. D., Smith, K. A., Follmer, S., Gerstenberg, T.
  2026; 165: 101814
- **Imagining and building wise machines: the centrality of AI metacognition.** *Trends in cognitive sciences*
  Johnson, S. G., Karimi, A. H., Bengio, Y., Chater, N., Gerstenberg, T., Larson, K., Levine, S., Mitchell, M., Rahwan, I., Schölkopf, B., Grossmann, I.
  2026
- **Corrigendum to 'If not me, then who? Responsibility and replacement' [Cognition 242 (2024) 105646].** *Cognition*
  Wu, S. A., Gerstenberg, T.
  2025: 106402
- **How children map causal verbs to different causes across development.** *Nature human behaviour*
  Rose, D., Zhang, S., Nichols, S., Markman, E. M., Gerstenberg, T.

2025

- **A signaling theory of self-handicapping.** *Cognition*
  Xiang, Y., Gershman, S. J., Gerstenberg, T.
  2025; 266: 106288
- **A framework for blaming willful ignorance.** *Current opinion in psychology*
  Kirfel, L., Gerstenberg, T., Zultan, R.
  2025; 65: 102090
- **Inference from social evaluation.** *Journal of personality and social psychology*
  Davis, Z. J., Allen, K. R., Kleiman-Weiner, M., Jara-Ettinger, J., Gerstenberg, T.
  2025
- **Causation, meaning, and communication.** *Psychological review*
  Beller, A., Gerstenberg, T.
  2025
- **From artifacts to human lives: Investigating the domain-generality of judgments about purposes.** *Journal of experimental psychology. General*
  Prinzing, M., Rose, D., Zhang, S., Tu, E., Concha, A., Rea, M., Schaffer, J., Gerstenberg, T., Knobe, J.
  2025
- **Causal-PIK: Causality-based Physical Reasoning with a Physics-Informed Kernel**
  Pares-Morlans, C., Yi, M., Chen, C., Wu, S. A., Antonova, R., Gerstenberg, T., Bohg, J.
  edited by Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2025: 44874-44886
- **Children use disagreement to infer what happened.** *Cognition*
  Amemiya, J., Heyman, G. D., Gerstenberg, T.
  2024; 250: 105836
- **Counterfactual simulation in causal cognition.** *Trends in cognitive sciences*
  Gerstenberg, T.
  2024
- **Anticipating the Risks and Benefits of Counterfactual World Simulation Models (Extended Abstract)**
  Kirfel, L., MacCoun, R., Icard, T., Gerstenberg, T., Association for the Advancement of Artificial Intelligence
  ASSOC COMPUTING MACHINERY.2024: 751
- **Self-Supervised Alignment with Mutual Information Learning to Follow Principles without Preference Labels**
  Franken, J., Zelikman, E., Rafailov, R., Gandhi, K., Gerstenberg, T., Goodman, N. D.
  edited by Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C.
  NEURAL INFORMATION PROCESSING SYSTEMS (NIPS).2024
- **MARPLE: A Benchmark for Long-Horizon Inference**
  Jin, E., Huang, Z., Franken, J., Liu, W., Cha, H., Brockbank, E., Wu, S., Zhang, R., Wu, J., Gerstenberg, T.
  edited by Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C.
  NEURAL INFORMATION PROCESSING SYSTEMS (NIPS).2024
- **Anticipating the Risks and Benefits of Counterfactual World Simulation Models (Extended Abstract)**
  Kirfel, L., MacCoun, R., Icard, T., Gerstenberg, T.
  edited by Das, S., Green, B. P., Varshney, K., Ganapini, M., Renda, A.
  ASSOC ADVANCEMENT ARTIFICIAL INTELLIGENCE.2024: 751
- **If not me, then who? Responsibility and replacement** *COGNITION*
  Wu, S. A., Gerstenberg, T.
  2024; 242
- **If not me, then who? Responsibility and replacement.** *Cognition*

Wu, S. A., Gerstenberg, T.
2023; 242: 105646

- **Making a positive difference: Criticality in groups.** *Cognition*
  Gerstenberg, T., Lagnado, D. A., Zultan, R.
  2023; 238: 105499
- **Mental jenga: A counterfactual simulation model of causal judgments about physical support.** *Journal of experimental psychology. General*
  Zhou, L., Smith, K. A., Tenenbaum, J. B., Gerstenberg, T.
  2023
- **Explanations Can Reduce Overreliance on AI Systems During Decision-Making** *PROCEEDINGS OF THE ACM ON HUMAN COMPUTER INTERACTION*
  Vasconcelos, H., Jorke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M., Krishna, R.
  2023; 7 (CSCW1)
- **Realism of Visual, Auditory, and Haptic Cues in Phenomenal Causality**
  Chase, E. D. Z., Gerstenberg, T., Follmer, S.
  edited by Bello, F., Choi, S., Kuchenbecker, K. J., MacLean, K.
  IEEE.2023: 306-312
- **MoCa: Measuring Human-Language Model Alignment on Causal and Moral Judgment Tasks**
  Nie, A., Zhang, Y., Amdekar, A., Piech, C., Hashimoto, T., Gerstenberg, T.
  edited by Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S.
  NEURAL INFORMATION PROCESSING SYSTEMS (NIPS).2023
- **Understanding Social Reasoning in Language Models with Language Models**
  Gandhi, K., Franken, J., Gerstenberg, T., Goodman, N. D.
  edited by Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S.
  NEURAL INFORMATION PROCESSING SYSTEMS (NIPS).2023
- **Active causal structure learning in continuous time.** *Cognitive psychology*
  Gong, T., Gerstenberg, T., Mayrhofer, R., Bramley, N. R.
  2022; 140: 101542
- **What would have happened? Counterfactuals, hypotheticals and causal judgements.** *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*
  Gerstenberg, T.
  2022; 377 (1866): 20210339
- **Do Humans Trust Advice More if it Comes from AI? An Analysis of Human-AI Interactions**
  Vodrahalli, K., Daneshjou, R., Gerstenberg, T., Zou, J., ACM
  ASSOC COMPUTING MACHINERY.2022: 763-777
- **Uncalibrated Models Can Improve Human-AI Collaboration**
  Vodrahalli, K., Gerstenberg, T., Zou, J.
  edited by Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A.
  NEURAL INFORMATION PROCESSING SYSTEMS (NIPS).2022
- **Inference From Explanation** *JOURNAL OF EXPERIMENTAL PSYCHOLOGY-GENERAL*
  Kirfel, L., Icard, T., Gerstenberg, T.
  2021
- **Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology.** *Cognition*
  Sosa, F. A., Ullman, T., Tenenbaum, J. B., Gershman, S. J., Gerstenberg, T.
  2021; 217: 104890
- **A counterfactual simulation model of causation by omission.** *Cognition*
  Gerstenberg, T., Stephan, S.
  2021; 216: 104842

- **Predicting responsibility judgments from dispositional inferences and causal attributions.** *Cognitive psychology*
  Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., Gerstenberg, T.
  2021; 129: 101412
- **A counterfactual simulation model of causal judgments for physical events.** *Psychological review*
  Gerstenberg, T., Goodman, N. D., Lagnado, D. A., Tenenbaum, J. B.
  2021
- **The Trajectory of Counterfactual Simulation in Development** *DEVELOPMENTAL PSYCHOLOGY*
  Kominsky, J. F., Gerstenberg, T., Pelz, M., Sheskin, M., Singmann, H., Schulz, L., Keil, F. C.
  2021; 57 (2): 253–68
- **A Causal Feeling: How Kinesthetic Haptics Affects Causal Perception**
  Chase, E. D. Z., Wolff, P., Gerstenberg, T., Follmer, S., IEEE
  IEEE.2021: 347
- **Expectations Affect Physical Causation Judgments** *JOURNAL OF EXPERIMENTAL PSYCHOLOGY-GENERAL*
  Gerstenberg, T., Icard, T.
  2020; 149 (3): 599–607
- **Moral Values Reveal the Causality Implicit in Verb Meaning.** *Cognitive science*
  Niemi, L. n., Hartshorne, J. n., Gerstenberg, T. n., Stanley, M. n., Young, L. n.
  2020; 44 (6): e12838
- **Causal Responsibility and Robust Causation.** *Frontiers in psychology*
  Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J. F., Usher, M.
  2020; 11: 1069
- **Quantitative causal selection patterns in token causation.** *PloS one*
  Morris, A. n., Phillips, J. n., Gerstenberg, T. n., Cushman, F. n.
  2019; 14 (8): e0219704
- **Time in Causal Structure Learning** *JOURNAL OF EXPERIMENTAL PSYCHOLOGY-LEARNING MEMORY AND COGNITION*
  Bramley, N. R., Gerstenberg, T., Mayrhofer, R., Lagnado, D. A.
  2018; 44 (12): 1880–1910
- **What's fair? How children assign reward to members of teams with differing causal structures** *COGNITION*
  Koskuba, K., Gerstenberg, T., Gordon, H., Lagnado, D., Schlottmann, A.
  2018; 177: 234–48
- **Lucky or clever? From expectations to responsibility judgments** *COGNITION*
  Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., Tenenbaum, J. B.
  2018; 177: 122–41
- **Eye-Tracking Causality.** *Psychological science*
  Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., Tenenbaum, J. B.
  2017; 28 (12): 1731-1744
- **Plans, Habits, and Theory of Mind.** *PloS one*
  Gershman, S. J., Gerstenberg, T., Baker, C. L., Cushman, F. A.
  2016; 11 (9): e0162246
- **Causal Conceptions in Social Explanation and Moral Evaluation: A Historical Tour** *PERSPECTIVES ON PSYCHOLOGICAL SCIENCE*
  Alicke, M. D., Mandel, D. R., Hilton, D. J., Gerstenberg, T., Lagnado, D. A.
  2015; 10 (6): 790–812
- **Causal superseding** *COGNITION*
  Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., Knobe, J.
  2015; 137: 196–209

- **Concepts in a Probabilistic Language of Thought** *CONCEPTUAL MIND: NEW DIRECTIONS IN THE STUDY OF CONCEPTS*
  Goodman, N. D., Tenenbaum, J. B., Gerstenberg, T.
  edited by Margolis, E., Laurence, S.
  2015: 623–53
- **Causal Responsibility and Counterfactuals** *COGNITIVE SCIENCE*
  Lagnado, D. A., Gerstenberg, T., Zultan, R.
  2013; 37 (6): 1036–73
- **When contributions make a difference: Explaining order effects in responsibility attribution** *PSYCHONOMIC BULLETIN & REVIEW*
  Gerstenberg, T., Lagnado, D. A.
  2012; 19 (4): 729–36
- **Spreading the blame: The allocation of responsibility amongst multiple agents** *COGNITION*
  Gerstenberg, T., Lagnado, D. A.
  2010; 115 (1): 166–71